DOI:10.16781/j.0258-879x.2018.08.0928

·综 述·

人工智能技术在基于电子病历研究中的应用与展望

唐仕超[1]，于观贞[2]，姜 磊[3*]
1. 上海交通大学医学院附属同仁医院风湿免疫科，上海 200336
2. 上海中医药大学附属龙华医院肿瘤七科，上海 200032
3. 海军军医大学（第二军医大学）长征医院风湿免疫科，上海 200003

[摘要] 人工智能技术在临床医学领域已取得突破性进展，如诊断、影像、疾病分期分级等。电子病历蕴含疾病描述、诊断、检查、治疗等大量临床数据，在医学专家和信息学家的共同参与下，利用人工智能技术挖掘电子病历数据的研究急剧增加。虽然该方法目前存在一些局限性，但与传统人工研究相比其具有更快速、经济、方便等优势，有望更好地服务于人类健康医学事业的发展。本文对利用人工智能技术挖掘电子病历数据的现状，包括相关技术、具体实例、局限性等进行综述。

[关键词] 人工智能；计算机化病案系统；数据源；自然语言处理
[中图分类号] R-37 **[文献标志码]** A **[文章编号]** 0258-879X(2018)08-0928-07

Application and prospect of artificial intelligence technology in electronic medical record research

TANG Shi-chao[1], YU Guan-zhen[2], JIANG Lei[3*]
1. Department of Rheumatology and Immunology, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200336, China
2. Department of Oncology (Ⅶ), Longhua Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai 200032, China
3. Department of Rheumatology and Immunology, Changzheng Hospital, Navy Medical University (Second Military Medical University), Shanghai 200003, China

[Abstract] Artificial intelligence technology has made breakthroughs in the field of clinical medicine, including diagnosis, imaging, and disease classification. Electronic medical record contains a large number of clinical data such as disease description, diagnosis, examination and treatment. With the participation of medical experts and information scientists, the studies of data mining of electronic medical record using artificial intelligence technology have greatly increased. Although now the method has some limitations, it is more rapid, economic and convenient compared with the traditional method, and is expected to promote the development of human health. In this paper, we reviewed the current status of data mining of electronic medical record using artificial intelligence technology, regarding related technologies, specific examples, and limitations.

[Key words] artificial intelligence; computed medical records systems; data source; natural language processing
[Acad J Sec Mil Med Univ, 2018, 39(8): 928-934]

人工智能（artificial intelligence）由被誉为“人工智能之父”的麦卡锡和明斯基于 1956 年首次提出。人工智能是研究和开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。我国人工智能经历了漫长而曲折的发展过程，直到改革开放后才得以蓬勃发展，2016 年蔡自兴院士[1]详细阐述了我国人工智能的发展史。人工智能属于计算机科学的一个分支，由机器学习、计算机视觉等不同领域组成，其最终目标是用智能机器模仿人类的思维活动和智力，并生产出一种新的、以与人类智能相似方式作出反应的智能系统，该领域研究包括机器人、语言识别、图像识别、自然语言处理和专家系统等。百度、腾讯、阿里巴巴、联想、科大讯飞等企业均已开展人工智

[收稿日期] 2018-06-21 **[接受日期]** 2018-07-10
[作者简介] 唐仕超，博士，副主任医师. E-mail: tagchaoyi@hotmail.com
*通信作者(Corresponding author). Tel: 021-66540109, E-mail: drlexjiang@163.com

能相关研究，其成果已在商业、农业、教育、金融、医疗、航天等行业发挥举足轻重的作用。

1 人工智能技术在医疗中的应用

IBM Watson 的发展是医疗领域的一个里程碑，目前在新药研发、辅助疾病诊断、辅助治疗、健康管理、医学影像、临床决策支持、医院管理、便携设备、康复医疗和生物医学研究等领域都有人工智能的实操案例。在我国，2017 年 3 月广州市发布战略性新兴产业蓝图：通过实施“IAB”计划，即发展新一代信息技术、人工智能、生物医药等战略性新兴产业，打造若干个千亿级产业集群。同年 7 月，国务院发布《关于新一代人工智能发展规划》，将人工智能提到国家战略的高度。中国数字医疗网（HC3i）发布的《2016－2017 年度人工智能＋医疗市场分析及趋势报告》显示，医疗人工智能的中国时代已经到来：2016 年中国人工智能＋医疗的市场规模已达到 96.61 亿元，2018 年有望达到 200 亿元。赵佳琦等[2]尝试运用骨骼肌超声图像计算机纹理分析技术研究肌肉损伤修复不同时期的超声影像表现，定量探讨了损伤及修复期声像图上肌束纹理特征变化，运用超声造影观察伤肢肌肉动脉期血管分布特征演变，并从组织病理学角度对照观察损伤及修复期的肌纤维结构和胶原分布特点、血管化程度及其排列分布。2018 年 4 月 12 日由海军军医大学（第二军医大学）长征医院影像科牵头举办的中国医学影像人工智能产学研用创新联盟成立大会暨首届医学影像人工智能高峰论坛中，既往人工需十几分钟才能从肺部影像中找出全部肺结节，而借助人工智能医疗影像辅助诊断只需 2 s 即得出结果，并标记出结节的位置、大小，其灵敏度达 99%，在小结节的识别上超过人类。自 2003 年人类基因组计划完成以来，单纯的全基因组关联研究（genome-wide association studies，GWAS）显著加速了对疾病遗传影响的研究[3]，二代测序也能确定罕见疾病的致病性单核苷酸多态性（single nucleotide polymorphism，SNP）[4]和一些常见疾病的重要调节因子[5-7]。Esteva 等[8]使用包含 129 450 张临床病理图像的数据集对中枢神经网络进行训练，并由 21 位皮肤科专家进行测试，最终验证了人工智能技术对皮肤癌的分类水平与皮肤科医师相当。Yu 等[9]采用人工智能技术识别肺结节病理切片，结果表明其可用于非小细胞肺癌的预后预测。 McCowan 等[10]将开发的软件用于肿瘤分期，该软件遵循 4 步过程来规范文本，可将其映射到统一医疗语言系统（unified medical language system，UMLS）以及检测和处理阴性结果。

2 电子病历（electronic medical record，EMR）为人工智能提供数据源

EMR 主要为临床护理而非研究设计，其还应用于临床诊疗、医疗保险和科学研究。EMR 含有丰富的信息，可提供患者临床诊疗的证据和临床研究表型的数据源。美国已建立的 EMR 库有电子病历与基因组学（electronic medical records and genomics，eMERGE）库、i2b2、PGPop、MVP 等，然而我国目前尚缺乏统一的 EMR 数据系统。此外，已成功建立的大型 EMR 库也存在局限性，如 EMR 数据本身不完善、记录偏倚、种族偏倚、大多只记录患病人群、不合语法文本、使用本地方言短语、使用缩写和拼写错误等；且由于各医疗中心应用的模板不同，导致患者在多个医疗中心的 EMR 数据不能导出，或不能与其他医疗中心的数据融合。虽然存在上述局限，EMR 丰富的资源仍决定人工智能技术应用于 EMR 的研究会发展迅速，应用前景广泛。

3 利用人工智能技术挖掘 EMR

利用 EMR 进行表型分型通常涉及跨学科协作。通常情况下，医学专家与信息学家一起创建并执行一种算法，以查询具有目标表型的受试者的 EMR 并随机选择病例进行回顾分析，在此过程中医学专家和信息学家都不可替代。医学专家了解表型及其在 EMR 中的表现，而信息学家熟知如何熟练运用人工智能技术分析并提取相应信息。验证是该过程的另一重要部分，不仅可以衡量算法的性能，还可以增强其在不同机构间的共享能力[11]。

从 20 世纪 90 年代开始，一些机构开始从志愿者身上采集 DNA 样本并将其存入生物样本库，利用人工智能技术与已经擦除识别信息的相应 EMR 关联，这些与 EMR 相关的 DNA 生物库有可能推动潜在临床表型遗传学的发现[12-13]。利用人工智能技术可以重复使用患者 EMR 中相关的遗传数据进行不同研究[14]，与传统人工研究相比，这大大

降低了关联研究的成本，减少了为每项研究表型招募患者的费用。研究发现基于 EMR 的人工智能技术可以降低每名受试者的研究费用高达 82%，且基于 EMR 的人工智能研究花费的时间短于传统人工研究[15-16]。

科研人员可以利用 EMR 研究不同表型，通过人工智能技术快速进行相关研究，如药物剂量反应[17-19]。目前通过人工或人工智能技术得出的相关表型，利用 eMERGE 和其他数据库的表型库可以在 Phenotype KnowledgeBase（*https://phekb.org/*）中查询，还可免费查询各种临床表型的记录。

4　利用人工智能技术挖掘 EMR 的技术支撑

目前生物医学信息学界已经开发出有据可查、易于分类的编码系统，如医学主题词表（medical subject headings，MeSH）、UMLS、国际疾病分类（international classification of diseases，ICD）、RxNorm、医学系统命名法-临床术语（systematized nomenclature of medicine-clinical terms，SNOMED-CT）、通用过程术语学（current procedural terminology，CPT）、观测指标标识符逻辑命名与编码系统（logical observation identifiers names and codes，LOINC）、解剖-治疗-化学代码（anatomical therapeutic chemical，ATC）等。EMR 包括一般信息、疾病描述、诊断、检查、治疗、生命体征、病程记录等，其中包含许多医学语言。基于上述标准化的编码系统，信息学家可以很容易提取出结构化的医学语言。首先利用医学语言提取和编码系统（medical language extraction and encoding system，MedLEE）、UMLS、MetaMap、HITex[20]、知识地图概念标识符（knowledge map concept identifier）等术语提取技术提取关键字；然后，基于临床经验和指南通过 RETE 算法、JBoss 规则引擎、特征选择算法、判别分析模型等自动生成规则的方法制定规则。

然而，对于非结构化的文本，信息学家不能应用标准化的医学语言提取关键字并形成规则，而需借助自然语言处理技术，如共指消解、时间分析、断言分类、语义网络技术、非结构化信息管理架构（unstructured information management architecture，UIMA）等。UIMA 是一个用于分析非结构化内容（文本、视频和音频）的组件架构和软件框架，其目的是为非结构化分析提供通用平台，从而提供能减少重复开发的可重用分析组件。临床文本分析和知识提取系统（clinical text analysis and knowledge extraction system，cTAKES）是通过使用 UIMA 和 OpenNLP 自然语言处理工具包构建的，它是临床自由文本的高级语义处理方法和模块的基础。支持向量机（support vector machine，SVM）是由 Corinna Cortes 和 Vapnik 于 1995 年首先提出，其在解决小样本、非线性及高维模式识别中表现出特有优势，并能推广应用到函数拟合等其他机器学习中。NegEx 是一种否定检测算法，其能检测非结构化文本中否定的临床表现，最终排除相关临床表现，避免出现假阳性；DEEPEN 是 Mehrabi 等[21]开发的一种否定算法，以减少 NegEx 的误报。

5　利用人工智能技术挖掘 EMR 的实例

2008 年 Wood 等[22]首次将 EMR 数据与 DNA 样本结合，他们对接受减肥手术患者进行了队列研究，收集 DNA 样本，从 EMR 中提取表型，并尝试复制了两种与冠心病和 2 型糖尿病相关的 SNP。

有研究人员基于 eMERGE 采用人工智能技术报道了欧洲裔美国人中与甲状腺功能减退症相关的叉头家族基因 *FOXE1* 的常见变异体[23]。Chen 等[24]利用 EMR 中的绝对淋巴细胞计数通过人工智能关联研究得出与衰老相关的 53 种基因。此外，还有学者采用人工智能技术利用 GWAS 中的基因和 EMR 中的红细胞沉降率[25]、红细胞计数[26]与水痘带状疱疹病毒感染[27]行关联研究。利用人工智能技术对 EMR 进行复杂疾病表型的定义已在诸多疾病中得到应用[28-33]。基于 EMR 得出的表型相比仅使用管理数据的准确性更高，因此其在临床和基因研究中应用更广泛[34-35]。Chase 等[36]使用自然语言处理技术挖掘门诊患者 EMR 中特定的症状体征，从而早期确定多发性硬化患者，缩短了诊断时间。此外，自然语言处理技术也可以应用于诊断肿瘤等初级保健机构经常遗漏的其他疾病，如类风湿关节炎[37]，最终实现早诊断和早治疗。Zheng 等[38]通过人工智能技术建立数据通知框架，基于 EMR 筛选 2 型糖尿病患者，发现该框架具有高性能识别；这种方法也可应用于

临床筛选其他有价值的样本，以最大限度降低选择偏倚。利用人工智能技术挖掘 EMR 还可应用于临床事件的预测[39-40]，既往已发表采用人工智能技术挖掘 EMR 建立近 30 种预测心力衰竭事件模型[41]。通过建立模型可以提高人们对该疾病的认识，从而探索微观机制，最终达到防病治病的目的。

6 利用人工智能技术挖掘 EMR 的局限性

我国缺乏 EMR 统一形式，其原因除 EMR 本身的局限性外，还可能是因为没有大型的样本库。迄今为止，大部分研究来自美国医院 EMR 研究组，可能与美国有完善的 EMR 库有关。基于 EMR 的研究数量有限且处于起步阶段，需要一套安全的去识别标准来保护患者的隐私，然而，去识别标准在信息学领域与医学领域不一致。Ford 等[42]指出信息学领域内的共识是报告测量精密度、召回率和 F-measure，而在医学领域通常是灵敏度和特异度。对于生物医学信息学研究人员，灵敏度等于召回率、正值预测值等于精密度，但生物医学领域以外的信息学并不使用特异度。为了更容易比较结果并从中得出结论，这两种文化必须更加融合。

EMR 包含大量临床信息，有诊断、化验结果等结构化数据和临床表现等非结构化数据。基于分类编码系统，结构化数据可以被快速检索[43]，但目前的分类编码系统各有优缺点[44]，没有一个编码体系能满足所有用户表型分析的需求。因此，对数据进行标准化是跨机构实现便携式表型解决方案的关键步骤。对于非结构化数据，其在自然语言处理下仍较难准确检索，且相关诊断可能只在临床记录中提及，需要不断完善算法。近年来，大量自然语言处理系统已被推荐用于从临床笔记中提取信息，现有许多公开可用的自然语言处理系统，如 cTAKES（Apache cTAKES，*http://ctakes.apache.org/*）、MedLEE[45]和 KMCI[46]。然而，由于所用语言的复杂性以及缺乏描述临床概念之间关系的明确语义资源，隐藏在笔记中的微妙关系仍难以提取[47-48]。有些研究需要将结构化与非结构化信息编织在一起以创建表型算法[49-50]。

EMR 可以用不同的方式记录和存储数据，如体质量可能会以不同的单位（千克、克和磅）记录和存储在 EMR 系统中，这可能导致错误的体质指数[36]。首字母缩略词可能有多种含义，如 RA 可以指类风湿性关节炎、右心房、房间空气和右臂，PD 可以指帕金森病和人格障碍，这种现象在临床 EMR 中经常见到[51]。这种不准确性通常不会误导临床医师对患者的诊断或治疗，因为临床医师可以根据上下文和医学知识辨别相关错误或解码缩写词，但计算机缺乏这种知识会使其难以检测准确的信息，从而导致误报。

目前可在 EMR 系统或机构中共享复杂表型的计算方法不存在，每个站点必须由本地信息学人员部署算法，并需要进行手动图表审核才能生效。如果某些表型的阳性预测值较低，可能需要对所有记录进行人工调整[17,52]。成功的表型分型可能需要临床医师、信息学家与其他领域专家共同合作开发验证算法。Wei 等[53]评估了 3 种主要 EMR 组分的表型分型，即 ICD 诊断代码、基本注释和特定药物，得出多个 EMR 组分相比单一组分可为表型提供更一致和更高的性能。Teixeira 等[54]评估 EMR 不同组分识别高血压个体时也得出同样结论。

7 展 望

从既往经验中可以发现，定义准确的表型已成为基于 EMR 遗传研究的关键步骤，该过程通常需要医学专家和信息学家共同参与不断迭代更新[55]。2009 年美国颁布的经济和临床健康信息技术法可能会增加 EMR 在基因研究中的可用性。由于相关规定旨在提高临床信息交换能力，大规模采用经过认证的 EMR 和达成的互操作性标准将加速表型与遗传数据在不同系统间的交换，从而形成更强大的 EMR 云[56]。然而，目前还没有适用于多个 EMR 系统和网站的标准可以使表型算法自动化、完全可计算、便捷执行，最接近的是质量数据模型[57]，然而目前该模型在表型算法中并不能应用于自然语言深度学习或复杂的方法[58]。

基于 EMR 的人工智能技术可提高传统基因的研究效率，充足的 EMR 数据有助于提取更可靠的表型。迄今为止，具有遗传数据的 EMR 生物样本库相对较小，但近期的预期是数百万 EMR 数据可用的患者可通过生物库获得基因数据。2015 年 1 月 20 日美国成立的精准医学项目将推动 EMR 生物样本库数据的进一步扩大。EMR 要适应时代的

需求，必须创建和采用新的标准，并且应该改进决策支持，以确保遗传学研究结果无缝集成到临床工作流程中。

目前 EMR 中的许多数据仍不可计算，进一步工作包括标准化 ICD-9-CM 、SNOMED-CT、RxNorm 等词汇表，而新知识和结构化医疗术语的应用可能会提高未来 EMR 的“可计算性”。通过这种方式，计算机可以推断出病毒性肺炎是一种感染性肺炎，它是病毒在肺部造成的病症，而不是单纯的病原体。

Shivade 等[59]对近年使用不同人工智能方法研究 EMR 的表型进行了比对，结合目前现状，并未得出哪种人工智能方法更适用于挖掘 EMR 的表型。正如 Deo 和 Nallamothu[60]所说，这不是一条容易的成功之路，但早期尝试有益于人们更了解人工智能，为后续研究奠定坚实的基础。

综上所述，人工智能技术已在临床医学的诸多领域取得显著成就，随着科技的进步及政策的扶持，相信未来研究会克服 EMR 的局限性，通过更先进的人工智能技术探索基因型与表型的关联，从基因水平治疗疾病；早期诊断相关疾病，实现早诊断和早治疗；筛选研究样本，提高研究的准确性和实用性；建立模型，提高对疾病的认识等，最终更好地服务于人类健康医学事业的发展。

[参 考 文 献]

[1] 蔡自兴. 中国人工智能 40 年[J]. 科技导报,2016,34:12-32.

[2] 赵佳琦,徐琪,章建全,黄禾菁,刁宗平. 骨骼肌超声诊断迈向人工智能新领域:计算机辅助骨骼肌损伤超声定量诊断[J]. 第二军医大学学报,2017,38:1217-1224.
ZHAO J Q, XU Q, ZHANG J Q, HUANG H J, DIAO Z P. Ultrasound diagnosis of skeletal muscle promoted by artificial intelligence: a quantitative evaluation of injured skeletal muscle by computer-aided ultrasonographic texture analysis[J]. Acad J Sec Mil Med Univ, 2017, 38: 1217-1224.

[3] MANOLIO T A. Genomewide association studies and assessment of the risk of disease[J]. N Engl J Med, 2010, 363: 166-176.

[4] BOYCOTT K M, VANSTONE M R, BULMAN D E, MacKENZIE A E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation[J]. Nat Rev Genet, 2013, 14: 681-691.

[5] SIGMA Type 2 Diabetes Consortium; WILLIAMS A L, JACOBS S B, MORENO-MACÍAS H, HUERTA-CHAGOYA A, CHURCHHOUSE C, MÁRQUEZ-LUNA C, et al. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico[J]. Nature, 2014, 506: 97-101.

[6] WILLER C J, SCHMIDT E M. SENGUPTA S, PELOSO G M, GUSTAFSSON S, KANONI S, et al. Discovery and refinement of loci associated with lipid levels[J]. Nat Genet, 2013, 45: 1274-1283.

[7] WEEKE P, MUHAMMAD R, DELANEY J T, SHAFFER C, MOSLEY J D, BLAIR M, et al. Whole-exome sequencing in familial atrial fibrillation[J]. Eur Heart J, 2014, 35: 2477-2483.

[8] ESTEVA A, KUPREL B, NOVOA R A, KO J, SWETTER S M, BLAU H M, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542: 115-118.

[9] YU K H, ZHANG C, BERRY G J, ALTMAN R B, RÉ C, RUBIN D L, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features[J/OL]. Nat Commun, 2016, 16: 12474. doi: 10.1038/ncomms12474.

[10] McCOWAN I A, MOORE D C, NGUYEN A N, BOWMAN R V, CLARKE B E, DUHIG E E, et al. Collection of cancer stage data by classifying free-text medical reports[J]. J Am Med Inform Assoc, 2007, 14: 736-745.

[11] NEWTON K M, PEISSIG P L, KHO A N, BIELINSKI S J, BERG R L, CHOUDHARY V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network[J/OL]. J Am Med Inform Assoc, 2013, 20: e147-e154. doi: 10.1136/amiajnl-2012-000896.

[12] HRIPCSAK G, ALBERS D J. Next-generation phenotyping of electronic health records[J]. J Am Med Inform Assoc, 2013, 20: 117-121.

[13] WILKE R A, XU H, DENNY J C, RODEN D M, KRAUSS R M, McCARTY C A, et al. The emerging role of electronic medical records in pharmacogenomics[J]. Clin Pharmacol Ther, 2011, 89: 379-386.

[14] KHO A N, PACHECO J A, PEISSIG P L, RASMUSSEN L, NEWTON K M, WESTON N, et al. Electronic medical records for genetic research: results of the eMERGE consortium[J/OL]. Sci Transl Med, 2011, 3: 79re1. doi: 10.1126/scitranslmed.3001807.

[15] BOWTON E, FIELD J R, WANG S, SCHILDCROUT J S, VAN DRIEST S L, DELANEY J T, et al. Biobanks and electronic medical records: enabling cost-effective research[J/OL]. Sci Transl Med, 2014, 6: 234cm3. doi: 10.1126/scitranslmed.3008604.

[16] NIH RePORTER [EB/OL]. [2018-06-19]. *http://projectreporter.nih.gov/reporter.cfm*.

[17] DELANEY J T, RAMIREZ A H, BOWTON E, PULLEY J M, BASFORD M A, SCHILDCROUT J S, et al.

Predicting clopidogrel response using DNA samples linked to an electronic health record[J]. Clin Pharmacol Ther, 2012, 91: 257-263.

[18] WEI W Q, FENG Q, JIANG L, WAITARA M S, IWUCHUKWU O F, RODEN D M, et al. Characterization of statin dose response in electronic medical records[J]. Clin Pharmacol Ther, 2014, 95: 331-338.

[19] RAMIREZ A H, SHI Y, SCHILDCROUT J S, DELANEY J T, XU H, OETJENS M T, et al. Predicting warfarin dosage in European-Americans and African-Americans using DNA samples linked to an electronic health record[J]. Pharmacogenomics, 2012, 13: 407-418.

[20] ZENG Q T, GORYACHEV S, WEISS S, SORDO M, MURPHY S N, LAZARUS R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system[J/OL]. BMC Med Inform Decis Mak, 2006, 6: 30. doi: 10.1186/1472-6947-6-30.

[21] MEHRABI S, KRISHNAN A, SOHN S, ROCH A M, SCHMIDT H, KESTERSON J, et al. DEEPEN: a negation detection system for clinical text incorporating dependency relation into NegEx[J]. J Biomed Inform, 2015, 54: 213-219.

[22] WOOD G C, STILL C D, CHU X, SUSEK M, ERDMAN R, HARTMAN C, et al. Association of chromosome 9p21 SNPs with cardiovascular phenotypes in morbid obesity using electronic health record data[J]. Genomic Med, 2008, 2(1/2): 33-43.

[23] DENNY J C, CRAWFORD D C, RITCHIE M D, BIELINSKI S J, BASFORD M A, BRADFORD Y, et al. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies[J]. Am J Hum Genet, 2011, 89: 529-542.

[24] CHEN D P, WEBER S C, CONSTANTINOU P S, FERRIS T A, LOWE H J, BUTTE A J. Novel integration of hospital electronic medical records and gene expression measurements to identify genetic markers of maturation[J]. Pac Symp Biocomput, 2008: 243-254.

[25] KULLO I J, DING K, SHAMEER K, McCARTY C A, JARVIK G P, DENNY J C, et al. Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate[J]. Am J Hum Genet, 2011, 89: 131-138.

[26] KULLO I J, DING K, JOUNI H, SMITH C Y, CHUTE C G. A genome-wide association study of red blood cell traits using the electronic medical record[J/OL]. PLoS One, 2010, 5: e13011. doi: 10.1371/journal.pone.0013011.

[27] CROSSLIN D R, CARRELL D S, BURT A, KIM D S, UNDERWOOD J G, HANNA D S, et al. Genetic variation in the HLA region is associated with susceptibility to herpes zoster[J]. Genes Immun, 2014, 16: 1-7.

[28] COLOMA P M, VALKHOFF V E, MAZZAGLIA G, NIELSSON M S, PEDERSEN L, MOLOKHIA M, et al. Identification of acute myocardial infarction from electronic healthcare records using different disease coding systems: a validation study in three European countries[J/OL]. BMJ Open, 2013, 3: e002862. doi: 10.1136/bmjopen-2013-002862.

[29] TU K, MITIKU T, GUO H, LEE D S, TU J V. Myocardial infarction and the validation of physician billing and hospitalization data using electronic medical records[J]. Chronic Dis Can, 2010, 30: 141-146.

[30] KOTTKE T E, BAECHLER C J. An algorithm that identifies coronary and heart failure events in the electronic health record[J/OL]. Prev Chronic Dis, 2013, 10: E29. doi: 10.5888/pcd10.120097.

[31] KHO A N, HAYES M G, RASMUSSEN-TORVIK L, PACHECO J A, THOMPSON W K, ARMSTRONG L L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study[J]. J Am Med Inform Assoc, 2012, 19: 212-218.

[32] CARROLL R J, THOMPSON W K, EYLER A E, MANDELIN A M, CAI T, ZINK R M, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records[J/OL]. J Am Med Inform Assoc, 2012, 19: e162-e169. doi: 10.1136/amiajnl-2011-000583.

[33] LIAO K P, CAI T, GAINER V, GORYACHEV S, ZENG-TREITLER Q, RAYCHAUDHURI S, et al. Electronic medical records for discovery research in rheumatoid arthritis[J]. Arthritis Care Res, 2010, 62: 1120-1127.

[34] KOHANE I S. Using electronic health records to drive discovery in disease genomics[J]. Nat Rev Genet, 2011, 12: 417-428.

[35] DENNY J C. Chapter 13: mining electronic health records in the genomics era[J/OL]. PLoS Comput Biol, 2012, 8: e1002823. doi: 10.1371/journal.pcbi.1002823.

[36] CHASE H S, MITRANI L R, LU G G, FULGIERI D J. Early recognition of multiple sclerosis using natural language processing of the electronic health record[J/OL]. BMC Med Inform Decis Mak, 2017, 17: 24. doi: 10.1186/s12911-017-0418-4.

[37] ZHOU S M, FERNANDEZ-GUTIERREZ F, KENNEDY J, COOKSEY R, ATKINSON M, DENAXAS S, et al. Defining disease phenotypes in primary care electronic health records by a machine learning approach: a case study in identifying rheumatoid arthritis[J/OL]. PLoS One, 2016, 11: e0154515. doi: 10.1371/journal.pone.0154515.

[38] ZHENG T, XIE W, XU L, HE X, ZHANG Y, YOU M, et

al. A machine learning-based framework to identify type 2 diabetes through electronic health records[J]. Int J Med Inform, 2017, 97: 120-127.

[39] CHOI E, BAHADORI M T, SCHUETZ A, STEWART W F, SUN J. Doctor AI: predicting clinical events via recurrent neural networks[J]. JMLR Workshop Conf Proc, 2016, 56: 301-318.

[40] RAJKOMAR A, OREN E, CHEN K, DAI A M, HAJAJ N, LIU P J, et al. Scalable and accurate deep learning for electronic health records[J/OL]. npj Dig Med, 2018, 1: 18. doi: 10.1038/s41746-018-0029-1.

[41] ECHOUFFO-TCHEUGUI J B, GREENE S J, PAPADIMITRIOU L, ZANNAD F, YANCY C W, GHEORGHIADE M, et al. Population risk prediction models for incident heart failure: a systematic review[J]. Circ Heart Fail, 2015, 8: 438-447.

[42] FORD E, CARROLL J A, SMITH H E, SCOTT D, CASSELL J A. Extracting information from the text of electronic medical records to improve case detection: a systematic review[J]. J Am Med Inform Assoc, 2016, 23: 1007-1015.

[43] TATE A R, BELOFF N, AL-RADWAN B, WICKSON J, PURI S, WILLIAMS T, et al. Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface[J]. J Am Med Inform Assoc, 2014, 21: 292-298.

[44] 秦宇辰,吴骋,王志勇,贺佳. 计算机辅助医疗信息标准化编码的现况及发展[J]. 中国数字医学,2018,13:9-12.

[45] SHORTLIFFE E H, CIMINO J J. Biomedical informatics: computer applications in health care and biomedicine[M/OL]. 4th ed. New York: Springer, 2014. *https://www.springer.com/us/book/9780387362786#otherversion=9780387289861*.

[46] DENNY J C, SMITHERS J D, MILLER R A, SPICKARD A 3rd. "Understanding" medical school curriculum content using KnowledgeMap[J]. J Am Med Inform Assoc, 2003, 10: 351-362.

[47] UZUNER Ö, SOUTH B R, SHEN S, DUVALL S L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text[J]. J Am Med Inform Assoc, 2011, 18: 552-556.

[48] NADKARNI P M, OHNO-MACHADO L, CHAPMAN W W. Natural language processing: an introduction[J]. J Am Med Inform Assoc, 2011, 18: 544-551.

[49] WEI W Q, LEIBSON C L, RANSOM J E, KHO A N, CARABALLO P J, CHAI H S. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughputclinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus[J]. J Am Med Inform Assoc, 2012, 19: 219-224.

[50] WEI W Q, TAO C, JIANG G, CHUTE C G. A high throughput semantic concept frequency based approach for patient identification: a case study using type 2 diabetes mellitus clinical notes[J]. AMIA Annu Symp Proc, 2010, 2010: 857-861.

[51] MOON S, PAKHOMOV S, LIU N, RYAN J O, MELTON G B. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources[J]. J Am Med Inform Assoc, 2014, 21: 299-307.

[52] OVERBY C L, PATHAK J, GOTTESMAN O, HAERIAN K, PEROTTE A, MURPHY S, et al. A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury[J/OL]. J Am Med Inform Assoc, 2013, 20(e2): e243-e252. doi: 10.1136/amiajnl-2013-001930.

[53] WEI W Q, TEIXEIRA P L, MO H, CRONIN R M, WARNER J L, DENNY J C. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance[J/OL]. J Am Med Inform Assoc, 2016, 23: e20-e27. doi: 10.1093/jamia/ocv130.

[54] TEIXEIRA P L, WEI W Q, CRONIN R M, MO H, VANHOUTEN J P, CARROLL R J, et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals[J]. J Am Med Inform Assoc, 2017, 24: 162-171.

[55] GURWITZ D, PIRMOHAMED M. Pharmacogenomics: the importance of accurate phenotypes[J]. Pharmacogenomics, 2010, 11: 469-470.

[56] MARSOLO K, SPOONER S A. Clinical genomics in the world of the electronic health record[J]. Genet Med, 2013, 15: 786-791.

[57] JIANG G, SOLBRIG H R, PATHAK J, CHUTE C G. Developing a standards-based information model for representing computable diagnostic criteria: a feasibility study of the NQF Quality Data Model[J]. Stud Health Technol Inform, 2015, 216: 1097.

[58] THOMPSON W K, RASMUSSEN L V, PACHECO J A, PEISSIG P L, DENNY J C, KHO A N, et al. An evaluation of the NQF Quality Data Model for representing Electronic Health Record driven phenotyping algorithms[J]. AMIA Annu Symp Proc, 2012, 2012: 911-920.

[59] SHIVADE C, RAGHAVAN P, FOSLER-LUSSIER E, EMBI P J, ELHADAD N, JOHNSON S B, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records[J]. J Am Med Inform Assoc, 2014, 21: 221-230.

[60] DEO R C, NALLAMOTHU B K. Learning about machine learning: the promise and pitfalls of big data and the electronic health record[J]. Circ Cardiovasc Qual Outcomes, 2016, 9: 618-620.

[本文编辑] 曾奇峰